Ikram Babs-Lawal
29/06/2025

---

1. Introduction

This report summarizes the approach, methods, and findings of my submission for the CuraJOY Impact Fellowship recruitment challenge. The goal was to build an accurate cyberbullying detection model and document the approach. Addressing cyberbullying is critical for the well-being of young people online, and this project aims to contribute a practical solution.

2. Dataset Overview

The dataset consists of text samples labeled as cyberbullying (`True`) or non-bullying (`False`). It was provided in a Python dictionary format and contained diverse examples of online communication. After loading, the data was explored to confirm class distribution and content diversity.

Key points:

- The dataset required cleaning and transformation before modeling.
- Class distribution analysis showed potential imbalance between bullying and non-bullying samples, a common issue in real-world scenarios.

3. Data Preprocessing

I applied the following steps to prepare the text data:

- Converted text to lowercase for consistency.
- Removed URLs and punctuation to reduce noise.
- Normalized whitespace.
- Used TF-IDF vectorization with 5,000 maximum features to transform text into numerical representations capturing word importance.

These preprocessing steps aimed to reduce irrelevant variability while preserving meaningful signals related to cyberbullying.

4. Model Development

A Logistic Regression model was chosen as a strong and interpretable baseline for binary text classification tasks. Key details:

- Data was split into an 80/20 train-test split, stratified by label to maintain class proportions.
- TF-IDF features were fitted on training data and transformed for both train and test sets.
- The Logistic Regression classifier was trained with default parameters but a higher max iteration count (1000) to ensure convergence.

The choice of Logistic Regression provides explainable outputs through feature weights and sets a benchmark for future, more complex models.

5. Model Evaluation

Model performance was evaluated on the test set using:

- Accuracy, precision, recall, and F1-score, focusing on minimizing false positives (flagging non-bullying content incorrectly) and false negatives (missing true bullying).
- A confusion matrix was plotted to visualize prediction outcomes.

Sample results:

- Precision: 0.87
- Recall: 0.99
- F1-score: 0.93

These metrics suggest the model is effective at distinguishing bullying content from normal conversation but highlight areas for improvement in handling ambiguous or sarcastic texts.

6. Research & Ethics Considerations

I designed a validation study framework to ethically and rigorously evaluate intervention effectiveness:

- Randomized control design comparing intervention vs. no intervention groups.
- Clear metrics such as reduction in harmful behavior and time to recurrence.
- Ethical considerations included informed consent, IRB approval for minors, privacy protection (anonymous IDs), and avoidance of interventions that could cause shame or distress.
- Bias mitigation strategies were included, such as stratified randomization and demographic audits.

7. Bias & Data Quality Analysis

Functions were implemented to:

- Detect potential demographic biases in model predictions (gender, race, age, platform).
- Assess data quality by checking for patterns indicating annotator fatigue or systematic disagreements in future annotation tasks.

These checks ensure fairness and high data quality, both crucial in sensitive tasks like cyberbullying detection.

8. Insights & Recommendations

Key takeaways and future improvements include:

- Handling sarcasm, slang, and code-switching is essential to reduce misclassifications.
- Advanced models like transformers (e.g., BERT) could improve understanding of nuanced language.

- Gathering more diverse data from varied platforms and age groups would improve generalization.

- Regularly auditing models and interventions is necessary to ensure ethical, fair, and effective operation.

9. Conclusion

Through data exploration, model development, evaluation, and ethical research design, this project demonstrates my commitment and ability to tackle complex real-world problems like cyberbullying detection. I look forward to the opportunity to contribute to CuraJOY's mission.